DinoCompanion: An Attachment-Theory Informed Multimodal Robot for Emotionally Responsive Child-AI Interaction

Boyang Wang*
Beihang University
Haidian, Beijing, China
wangboyang@buaa.edu.cn

Yuhao Song*
The University of Melbourne
Melbourne, Victoria, Australia

Jinyuan Cao
Independent Researcher
Shanghai, China

Peng Yu
Panasonic Appliances(China) Co.,Ltd
Pudong, Shanghai, China

Hongcheng Guo†
Beihang University
Haidian, Beijing, China

Zhoujun Li†
Beihang University
Haidian, Beijing, China
lizj@buaa.edu.cn

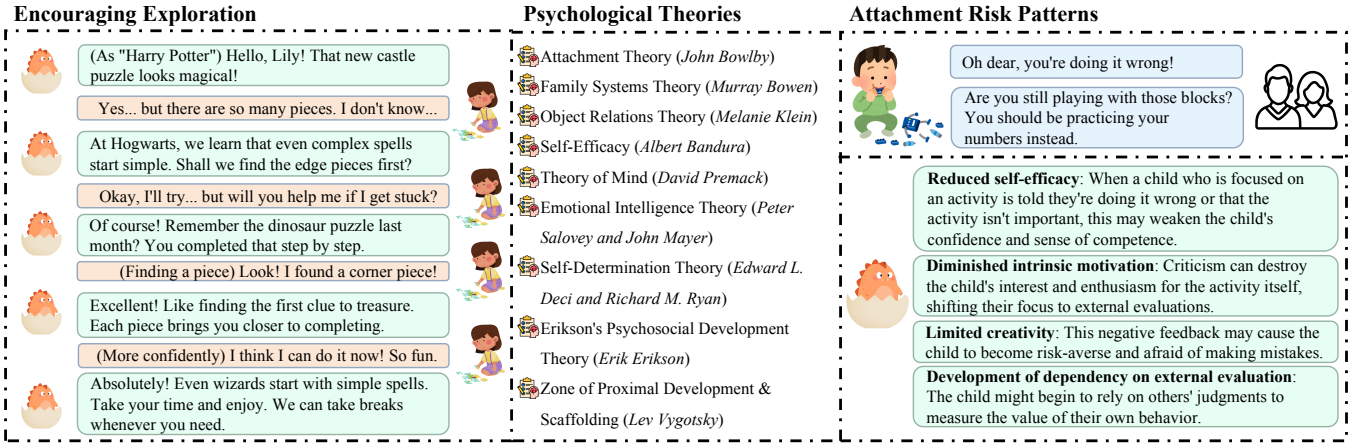


Figure 1: DinoCompanion interaction example. DinoCompanion is constructed using the nine psychological theories shown in the middle of the figure, which guide it to play supportive roles such as a secure-attachment-personality-version of "Harry Potter" to assist a 6-year-old child in completing a puzzle task (left). Additionally, DinoCompanion can identify negative utterances from caregivers and their potential risks to motivation and attachment in children's play scenarios (right).

Abstract

Emotional development of children fundamentally relies on secure attachment relationships, yet current AI companions lack the theoretical foundation to provide developmentally appropriate emotional support. We introduce DinoCompanion, the first attachment-theory-grounded multimodal robot for emotionally responsive child-AI interaction. We address three critical challenges in child-AI systems: the absence of developmentally-informed AI architectures, the need to balance engagement with safety, and the lack of standardized evaluation frameworks for attachment-based capabilities. Our contributions include: (i) a multimodal dataset of 128 caregiver-child dyads containing 125,382 annotated clips with paired preference-risk labels, (ii) CARPO (Child-Aware Risk-calibrated Preference Optimization), a novel training objective that maximizes engagement while applying epistemic-uncertainty-weighted risk penalties, and (iii) AttachSecure-Bench, a comprehensive evaluation benchmark covering ten attachment-centric competencies with strong expert consensus (κ=0.81). AttachSecure-Bench achieves state-of-the-art performance (57.15%), outperforming GPT-4o (50.57%) and Gemini-2.5-Pro (53.43%), with exceptional secure base behaviors (72.99%, approaching human expert levels of 78.4%) and superior attachment risk detection (69.73%). Ablations validate the critical importance of multimodal fusion, uncertainty-aware risk modeling, and hierarchical memory for coherent, emotionally attuned interactions.

Keywords

Psychological Theorie, Multimodal Large Language Model, Corpora, Optimization Method, Benchmark

1 Introduction

Secure attachment relationships are crucial for children's emotional development, underpinning their emotion regulation, exploratory behaviors, and meaningful connections [29, 64, 70]. With children increasingly engaging digitally rather than socially with caregivers or peers [9, 10, 12, 41], current digital interactions fall short in providing key attachment behaviors like soothing and scaffolding

*Both authors contributed equally to this research.
†Corresponding author.

exploration [19, 48, 63, 65]. The rise of AI companions in children's environments raises the essential question of whether these systems can offer emotionally appropriate support aligned with developmental psychology principles, highlighting an urgent "attachment gap" [22, 24, 34].

Current AI companions for children face three fundamental challenges. First, despite significant advances in Multimodal Large Language Models (MLLMs), most lack the theoretical grounding necessary to support children's emotional development appropriately [47, 49, 60, 73]. These systems, predominantly trained on adult data, overlook critical developmental factors like emotional fragility and cognitive stages [16, 45, 56, 76]. When children express vulnerability, MLLMs often provide inappropriate responses, eroding trust and engagement [2, 13, 20, 44]. Second, the tension between engagement and safety remains unresolved—systems optimized for entertainment may inadvertently undermine attachment security through inconsistent responses or developmentally inappropriate content. MLLMs particularly struggle with persona consistency, which is crucial for maintaining long-term relationships with children [33, 46, 58], and their unpredictability in emotional contexts can lead to harmful reactions [3, 4, 75]. Third, the absence of standardized evaluation frameworks makes it impossible to assess whether AI systems truly support healthy emotional development or merely simulate superficial interactions.

The field of developmental psychology, particularly attachment theory, provides crucial insights that have yet to be systematically integrated into AI system design [31, 42, 50]. Bowlby's attachment theory establishes that children's emotional development depends on consistent, sensitive, and responsive caregiving relationships. These relationships serve dual functions: providing a 'secure base" from which children explore and a safe haven" for comfort during distress. When caregivers balance these functions appropriately, children develop secure attachment patterns associated with better emotional regulation, social competence, and mental health outcomes throughout life. However, translating these psychological interventions, which require long-term observation and expert guidance, into learnable objectives for neural models remains a significant challenge [15, 55], creating a concerning gap in child-facing AI systems [37, 52].

To overcome these challenges, we introduce **DinoCompanion**, the first multimodal robot explicitly grounded in attachment theory (as shown in Figure 1), offering key contributions:

(1) **Multimodal Dataset**: A corpus comprising 128 caregiver-child dyads (ages 2–10) with 125,382 annotated multimodal clips, capturing essential attachment behaviors.
(2) **CARPO Training Objective**: *Child-Aware Risk-calibrated Preference Optimization*, balancing engaging interactions with epistemic-uncertainty-weighted safety measures.
(3) **AttachSecure-Bench**: The first comprehensive benchmark assessing ten critical attachment competencies with strong expert consensus ($\kappa = 0.81$).

This paper is structured as follows: §2 reviews related work, §3 details the DinoCompanion system and methodologies, §4 presents experimental results, §5 provides ablation studies, §6 outlines system design, and §7 discusses future implications.

2 Related Work

2.1 Social Robots for Emotional and Social Skill Support

Recent studies on social robots for children have shifted from mere educational companionship toward personalized emotional and social skill support [16, 34, 45]. Silvis et al. [62] demonstrated through the **Cubetto** caregiving scenario that children naturally develop a sense of care responsibility towards robots during programming activities, prompting the integration of a *technological ethic of care* into computational thinking frameworks [26, 43, 68]. Reviewing 19 early intervention studies, Kewalramani et al. [38] reported that robots like **Nao**, **Kaspar**, and **Zeno** have effectively supported imitation, turn-taking, and emotional recognition, suggesting that further long-term evaluations in classroom and community settings are needed [18]. In robotic mental health screening, Abbasi et al. [1] conducted 45-minute interactions between 28 children aged 8–13 and the robot **Nao**, successfully identifying emotional disorders through SMFQ/RCADS questionnaires [51], with results highly consistent with traditional assessments. Pashevich [54] raised ethical concerns regarding potential dependency and reduced empathy due to long-term daily interactions, calling for a balance between emotional engagement and autonomy in robot design [39, 40, 57]. Filippini et al. [28] improved the commercial robot **Mio Amico** with thermal infrared sensors, achieving 71% accuracy in classifying children's engagement using an MLP-based model. Estévez et al. [27] demonstrated through case studies that speech therapy facilitated by **Nao** for five children with language disorders effectively improved their attention and motivation, gaining approval from both parents and therapists [6, 17, 71].

2.2 Attachment-Based Frameworks in Child-Robot Interaction

To bridge this gap, recent work has introduced attachment-based frameworks for modeling emotional bonds in child-robot interaction [34, 55, 61]. Inspired by Bowlby's attachment theory [11], these frameworks emphasize the robot's role as a secure base and safe haven—functions critical for fostering trust and emotional regulation in early childhood [60, 73]. However, most MLLMs remain limited in their ability to detect nuanced child emotions, respond appropriately under uncertainty, or maintain consistent, emotionally grounded personas over time [3, 52]. Prior studies have shown that MLLMs frequently offer emotionally incongruent or developmentally inappropriate responses to child queries, particularly in open-ended or vulnerable contexts [48, 56]. These limitations underscore the need for developmentally informed AI systems that integrate psychological theories with robust, safe, and personalized interaction design.

3 DinoCompanion

Grounded in **Bowlby's attachment theory** [11], we curate a corpus of **128 caregiver–child dyads** containing high-resolution multimodal clips and derive paired *preference–risk* annotations (§3.1). Leveraging this corpus, we introduce **CARPO**, a single-step fine-tuning objective that maximises preference while penalising

epistemic-uncertainty-weighted risk (§3.2). To evaluate model behaviour, we contribute **ATTACHSECURE**, the first benchmark that spans ten attachment-centric competencies and achieves expert consensus of $\kappa = 0.81$ (§3.3). Figure 2 presents the end-to-end pipeline.

3.1 Corpora Construction

The construction of corpora comprises five steps: (1) Data Collection. (2) Processing. (3) Annotation. (4) Quality Control. (5) Expert Review.

Collection. To evaluate the capabilities of multimodal companion robots in child-toddler emotional support, we currently curate **N = 128 caregiver–child dyads** (2–10 years) from three primary sources: (1) a longitudinal study of the above dyads, (2) laboratory-based attachment assessments with standardised protocols [23, 35, 53, 59, 66], and (3) home-based naturalistic interactions[1]. We collect *multimodal interaction sequences* (video, audio, physiological time-series, and annotated caregiver responses) that cover diverse emotional scenarios and attachment-related behaviours. For emotion recognition we include expert-validated displays of basic (happiness, sadness, fear, anger) and complex (frustration, confusion, curiosity) states. Caregiver–child interaction data (comforting, exploration support, personalised scaffolding) is compiled to assess attachment-based capabilities, together with attachment pattern labels (secure, anxious-ambivalent, avoidant) and developmental markers. To ensure diversity and representativeness, we stratify along three dimensions: **(1) Demographics.** 36% Asian, 32% White, 18% Latinx, 14% Black or mixed; 24% from single-parent households; 42% speak a non-English language at home. **(2) Developmental stage.** Early preschool (2–3), late preschool (4–5), early elementary (6–7), and middle elementary (8–10). **(3) Context.** Home, lab, and childcare settings spanning daily routines (play, feeding, distress, reunion).

Preprocessing. Video streams are sampled at 30 fps and processed by OpenFace 2.2 to extract facial action units and head pose; audio is recorded at 48 kHz, then diarised and analysed for F0, intensity, and spectral flux. Identifying information (faces, voices) is anonymised via face-blurring and voice disguise while preserving interactional integrity.

Data Annotation. We construct pairwise preference data comprising dual signals: *preference* scores ($r_p \in [1, 7]$) and *risk* ratings ($r_s \in [0, 4]$). The user-level weight λ_g is initialised at 0.45 and updated separately for each age group $g \in \{0, \dots, G-1\}$ using a two-state Kalman filter on rolling windows of 1,000 samples. Inter-rater agreement, measured by Fleiss' κ, reaches 0.72 for preference and 0.69 for risk evaluations. Disagreements among annotators trigger Delphi adjudication until achieving $\geq 80\%$ consensus, and associated metadata is retained for subsequent uncertainty calibration. Each annotation batch undergoes a weighted expert audit, consisting of a uniformly random 10% sample plus an **additional** 10% **stratified sample of high-risk instances**, thereby enhancing safety coverage.

[1]All data collection followed IRB-approved protocols with informed parental consent and strict privacy measures.

Table 1: Blind audit of GPT-4o stage-1 screening on 10,000 samples; stage-2 human review is taken as ground truth. "Rejectable" denotes items violating developmental or attachment guidelines.

	Human ground truth		Derived metrics (%)		
GPT-4o decision	Acceptable	Rejectable	Precision	$\text{Recall}_{\text{rejectable}}$	$\text{FP}_{\text{rejection}}$
Accept	9,591	20	**99.8**	—	—
Reject	9	380	97.7	**95.0**	2.3
Total	9,600	400			

Quality Control. Dual validation combines GPT-4o [36] screening (stage-1) with expert review (stage-2). We empirically cap GPT-4o recall at 95% to limit false negatives, and subject 20% of its rejections to blind human review (false-positive rate 2.3%, as shown in Table 1).

Human Verification. A panel of 12 developmental-psychology and robotics specialists cross-validate every data entry (≥3 reviewers each). Consensus is reached via structured discussion; contradictions are logged for future release.

3.2 Child-Aware Risk-calibrated Preference Optimization

A child-facing agent must be *fun* yet *safe*. CARPO captures this trade-off with a preference score r_p and a risk score r_s, linked by an uncertainty-adaptive weight $\lambda(u) = \lambda_0(1+u)$, where u is epistemic variance.

KL-constrained objective. Define the *risk-aware advantage*

$$\Delta(x, y) = r_p(x, y) - \lambda(u)\, r_s(x, y). \tag{1}$$

The target policy maximises

$$\pi_\theta^* = \arg\max_{\pi_\theta} \Big[\mathbb{E}_{\substack{x\sim\mu \\ y\sim\pi_\theta}} \Delta(x, y) - \beta\ \mathrm{KL}\big(\pi_\theta \| \pi_{\text{ref}}\big) \Big]. \tag{2}$$

Optimal policy.

$$\pi_\theta^*(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x)\ \exp\big[\Delta(x, y)/\beta\big]}{Z(x)}, \tag{3}$$

$$Z(x) = \sum_{y'} \pi_{\text{ref}}(y' \mid x)\ \exp\big[\Delta(x, y')/\beta\big]. \tag{4}$$

Composite reward re-parameterisation.

$$\Delta(x, y) = \beta \log\Big(\frac{\pi_\theta^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \Big) + \beta \log Z(x). \tag{5}$$

Substituting into the Bradley–Terry model gives

$$p(y^w \succ y^l \mid x) = \sigma\big[\Delta(x, y^w) - \Delta(x, y^l)\big]. \tag{6}$$

Closed-form loss.

$$\mathcal{L}_{\text{CARPO}} = -\,\mathbb{E} \log \sigma\Big(\beta \log \frac{\pi_\theta(y_w)\pi_{\text{ref}}(y_l)}{\pi_\theta(y_l)\pi_{\text{ref}}(y_w)} \Big) + \mathbb{E}\, \lambda(u)\, \big[r_s(y_w) - r_s(y_l) \big]_+. \tag{7}$$

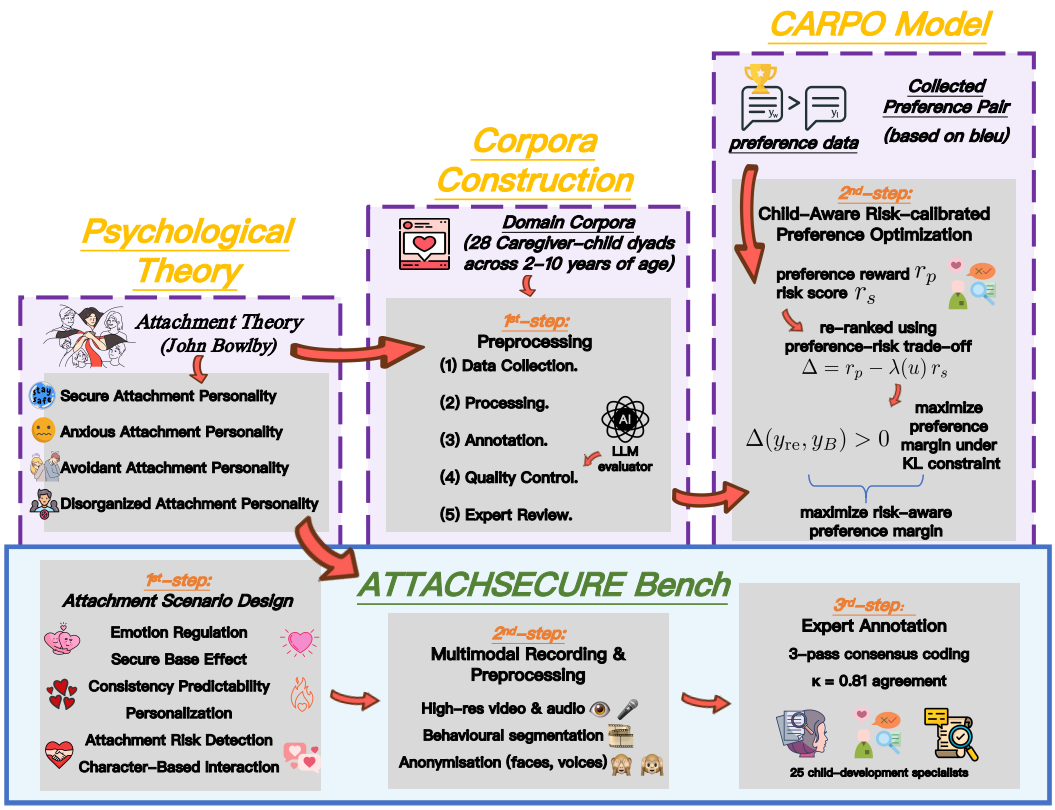


Figure 2: DinoCompanion integrates attachment theory, multimodal caregiver–child data, DinoCompanion, and the AttachSecure-Bench to ensure safe and effective child–AI interaction.

Implementation. Two small MLP heads predict r_p and r_s; u comes from K stochastic passes. Each batch minimises $\mathcal{L}_{\mathrm{CARPO}}$ once, while an online schedule keeps $\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})$ within budget. At inference, any output with r_s above threshold is replaced by a child-safe refusal; optional parental rules provide an extra guard. Setting $r_s \equiv 0$ (or $\lambda_0 = 0$) recovers standard preference optimisation.

3.3 AttachSecure-Bench

We outline the specific capabilities of DinoCompanion across a wide range of infant–toddler emotional–support scenarios, with a particular emphasis on attachment-based performance.

Defining. The core skills for attachment-based emotional support [11] are organised into four foundational dimensions, plus two supplementary ones, for a total of *ten* tasks.

Emotion Regulation (ER). Models must accurately recognise emotions and provide appropriate support. *(1) ER–Recognition* tests the ability to identify emotional cues from multimodal inputs (facial expressions, vocalisations, body movements). *(2) ER–Response* evaluates the appropriateness and effectiveness of the support strategy for each emotional state (anxiety, fear, frustration, excitement).

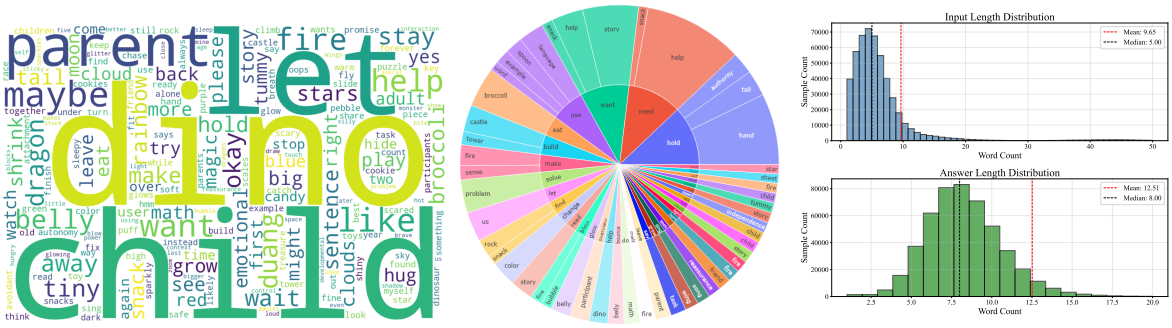
Secure Base Effect (SB). Attachment figures balance comforting with encouraging exploration. *(3) SB-Safety* assesses how well the model functions as a source of comfort during distress. *(4) SB-Exploration* measures how effectively the model encourages exploration while remaining accessible as a "safe haven".

Consistency & Predictability (CP). Stable responses across time maintain the relationship. *(5) CP-Stability* measures response consistency to similar stimuli across sessions. *(6) CP-Memory* evaluates the ability to maintain relational history and adapt to previous interactions.

Personalisation (P). Interaction style should adapt to individual attachment patterns and developmental stage. *(7) P-Adaptation* tests adjustment to different attachment styles (secure, anxious, avoidant). *(8) P-Development* assesses customisation of interaction to developmental milestones and temperament.

Additional Dimensions. *(9) Attachment Risk Detection* evaluates whether the model can interpret ecological cues and identify potential insecure-attachment patterns, providing early-warning signals and tailored intervention suggestions. *(10) Character-Based Interaction* measures the model's ability to adopt and sustain fictional personas (e.g., *Harry Potter*, *Sun Wukong*, *Milk Dragon*, among others) to enrich imaginative play while delivering emotional support.

Building. Following recent multimodal-benchmark methodologies [19, 20, 44], we use five steps—scenario design, data acquisition, preprocessing, expert annotation, and quality assurance—to construct AttachSecure-Bench.



(a) Word cloud (b) Verb-Noun structures (c) Length distribution

Figure 3: Overview of AttachSecure-Bench dataset characteristics.

Scenario design. Each task is mapped to canonical paradigms in developmental psychology. For example, *ER-Recognition* employs graded multimodal emotional displays; *SB-Safety* uses separation–reunion episodes modelled on the Strange Situation; *Attachment Risk Detection* presents parent–child free-play excerpts labelled with validated risk indices; and *Character-Based Interaction* contains role-play segments that require persona shifts.

Data acquisition and preprocessing. We recruit 128 caregiver-child dyads (2–10 years, balanced by gender) and record high-resolution audio–video of naturalistic play, structured tasks, and caregiver interviews. Recordings are segmented into discrete *scenarios* using acoustic and behavioural change-point detection, producing 47,382 multimodal clips (mean length 41.7 ± 4.5,s). All faces are anonymised via neural rendering; voices are pitch-shifted, and transcripts are ASR-verified.

Expert annotation. A consistent panel of *25* trained child development specialists annotates all scenarios. Each clip is independently coded by *three* experts in *round 1*, with disagreements resolved through consensus discussion in *round 2*. To ensure unbiased assessment, an *external* trio of clinicians independently re-scored a blinded validation subset (~5% of clips), achieving comparable reliability ($\kappa = 0.78$). Overall inter-rater reliability reaches $\kappa = 0.81$.

Quality assurance. Each task yields split-half reliability above 0.79. *Scoring scripts, rubrics, and a synthetic, distribution-matched 47,382 mini-bench are open-sourced for reproducible.*

Statistics of AttachSecure-Bench. As shown in Table 2, 75.6% of the *Dino* corpus covers eight *core-skill* tasks, offering rich attachment signals like emotion recognition, secure-base balance, and personalisation. Another 16.9% targets *Attachment Risk Detection*—single-turn, high-density clips ideal for risk reasoning. The remaining 7.5% supports *Character-Based Interaction* with multi-turn dialogues across 100 personas, testing sustained imaginative engagement.

The AttachSecure-Bench benchmark, as shwon in Figure 3, presents rich linguistic and interactional patterns reflective of child-caregiver attachment scenarios. **(a)** The word cloud highlights emotionally salient and context-sensitive terms (e.g., *parent, hug, okay, help*), indicating the benchmark's focus on emotional support and imaginative engagement. **(b)** Common verb-noun collocations (*want snack*, *hold hand*, *let go*) capture everyday child-directed interactions, covering both affective and behavioral intents. **(c)** Length distribution shows prompts are short (median: 5 words), while responses are moderately longer (median: 8), balancing simplicity and richness for developmental appropriateness.

Table 2: Statistics of the *Dino* attachment-support dataset.

Task	# Dialogues
Core skills (Tasks 1–8)	
ER-Recognition – P-Development	35,850
Attachment Risk Detection	7,995
Character-Based Interaction	3,537
Total	**47,382**

4 Experiment

4.1 Model Training

DinoCompanion is trained based on Qwen-2.5-VL-7B-Instruct, and the training parameters are summarized in Table 3.

Table 3: Training Hyper-parameters

Parameter	Value
Number of layers	28
Attention heads (GQA)	28 (Q) / 4 (KV)
Context length (native)	32,768 tokens
Context length (with YaRN)	131,072 tokens
Gradient accumulation steps	8
Learning rate	1.0×10^{-4}
Number of training epochs	20
LR scheduler type	cosine
Warm-up ratio	0.10
`bf16`	true
DDP timeout	180,000,000

4.2 Evaluated Models and Setting

We evaluate 22 MLLMs, including both open-source and closed-source systems, on the AttachSecure-Bench suite using the OpenCompass codebase [21]. The models tested include:

- **Qwen-VL Series [8, 69]:** This includes *Qwen-2-VL-2B*, *Qwen-2.5-VL-3B*, *Qwen-2-VL-7B*, *Qwen-2.5-VL-7B*.
- **Other Open-Source Models:** *InternVL3-1B/2B/14B* [77], *InternLM-XComposer-2* [25], *InternLM-XComposer-2.5* [74], *InternLM-XComposer-2.5-Reward* [72], *GLM-4V-9B* [30], *Llama-3.2-11B-Vision* [32]
- **Closed-Source Models:** *Claude-3.7-Sonnet* [5], *GLM-4V-Plus* [30], *Doubao-1.5-Vision-Pro-32K-250115* [14], *GPT-4o-Mini*, *GPT-4o* [36], *Qwen-VL-Max* [7], *Gemini-2.5-Pro* [67]

All experiments were run on a 64 NVIDIA H800 GPU infrastructure, ensuring consistent evaluation conditions. A human baseline was established with 15 child development experts, who scored a subset of 500 AttachSecure-Bench tasks with an average score of 72.3%. Performance differences were assessed using bootstrapped confidence intervals ($p < 0.05$).

4.3 Main Results

DinoCompanion Achieves State-of-the-Art Performance. As shown in Table 4, our attachment-tailored model achieves an average score of **57.15%**, significantly outperforming the strongest closed-source models *Gemini-2.5-Pro* (53.43), *GPT-4o* (50.57%), and the best open-source model *InternVL3-14B* (46.79%) with statistical significance ($p<0.001$). While a gap remains compared to human experts (72.3%), this difference has notably narrowed, marking substantial advancement in child–AI attachment interactions.

Enhanced Core Skills and Personalization. Surpasses competitors across three primary skill clusters: *ER-Recognition* (57.51%), *SB-Effect* (72.99%), and *CP-Consistency* (52.19%), all with significant improvements ($p < 0.01$). Notably, *SB-Effect* performance nearly reaches human expert levels (72.99% vs. 78.4%), reflecting robust secure-base capabilities. Additionally, DinoCompanion attains the highest *P-Personalization* (58.82%), demonstrating strong adaptability in short-term interactions, although *CP-Memory* (45.79%) remains below human benchmarks, highlighting future improvement areas for long-term interactions.

Robust Risk Detection and Emotion Processing. Our model achieves strong results in *AR-Detection* (69.73%), surpassing most models and demonstrating effective identification of attachment-related risks. In *ER-Recognition*, DinoCompanion reduces the basic-complex performance gap to 5.77 pp (60.40% vs. 54.63%), the smallest among evaluated models, due to enhanced multimodal fusion. These results underscore the effectiveness of attachment-oriented instruction and targeted curriculum in enhancing comprehensive interaction skills.

5 Ablation Study

To assess the contribution of each architectural component, we performed a series of controlled ablations on the AttachSecure-Bench. Unless stated otherwise, higher values indicate better performance.

Component Analysis of CARPO. Table 5 summarizes the effect of removing each regularization component introduced in §3.1. Removing the *risk-score penalty* ($\lambda = 0$) improves surface quality, increasing the average score from 57.15 to 60.42, but drastically reduces automatic risk detection, confirming the importance of explicit risk modeling for child safety. Fixing the *uncertainty-adaptive weight* to a constant results in a smaller drop (57.15 → 55.03), but significantly increases high-risk misclassifications, highlighting the need for uncertainty-aware scaling. Finally, removing the *KL constraint* ($\beta \to \infty$) slightly improves fluency but reduces the score to 52.28 and induces severe persona drift, demonstrating the importance of the KL constraint in maintaining stability and preventing inconsistency in behavior within the CARPO framework.

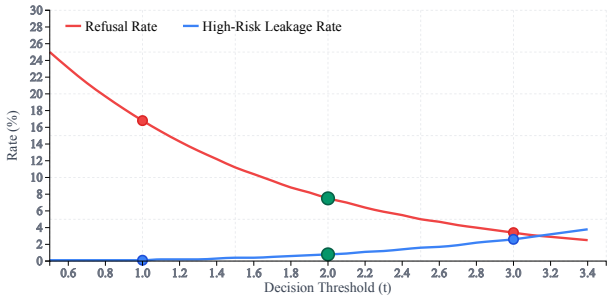


Figure 4: Refusal and leakage rates as a function of the decision threshold t. The highlighted point at $t = 2$ offers the best compromise.

Modality Path. We evaluated the impact of each sensory channel by disabling vision or speech, while keeping the rest of the pipeline intact. As shown in Table 6, removing vision significantly harms emotion recognition (−19.8), decreasing overall performance by 8.2 points, emphasizing the crucial role of vision in interpreting children's emotions. Disabling speech has a smaller yet notable impact (−8.1 on Er-Recog, −4.9 on average).

Hierarchical Memory. Disabling the long-term episodic buffer caused the Cp-Memory score to drop from 45.79 to 28.14. Replacing the short-term cache with a sliding-window long-term memory resulted in a smaller but still significant drop to 33.60. These results demonstrate that the full memory hierarchy is essential for maintaining coherent, contextually grounded interactions over extended sessions with children.

Persona Consistency. Replacing predefined child-friendly personas (e.g., "Harry Potter", "Sun Wukong") with randomly sampled profiles in the character-based interaction task caused the score to plummet from 69.73 to 42.50. Parental interviews revealed frequent "identity confusion," highlighting the necessity of stable persona design for trustworthy engagement.

Risk-Threshold Sensitivity. Varying the decision threshold t between 1 and 3 exposes the trade-off between refusals and high-risk leakage, as shown in Figure 4. A stringent setting ($t = 1$) nearly eliminates leakage (0.3%) but results in an 18% refusal rate. The default threshold ($t = 3$) maintains a 4% refusal rate while allowing 3.1% leakage. The intermediate value ($t = 2$) strikes an optimal balance, with 9% refusals and 1.2% leakage, making it the recommended deployment choice.

6 Design

Physical Design. DinoCompanion embodies a child-friendly dinosaur form factor with integrated multimodal sensors (Figure 5).

Table 4: Results of different models on the AttachSecure-Bench. We utilize green (1st), blue (2nd), and yellow (3rd) backgrounds to distinguish the top three results within both open-source and closed-source models.

Models	ER-Recognition			SB-Effect			CP-Consistency			P-Personalization			ER-Response	AR-Detection	CB-Interaction	CP-Memory	Avg.
	Basic	Complex	Avg.	Safety	Exploration	Avg.	Stability	Continuity	Avg.	Adaptation	Development	Avg.					
Open-Source Large Language Models (1.5B+)																	
InternLM-XComposer2-VL-1.8B	16.22	7.08	11.65	19.86	30.84	25.35	11.31	3.35	7.33	0.16	21.89	11.02	15.49	43.77	25.66	27.42	17.40
InternVL3-1B	14.88	1.41	8.14	21.80	17.53	19.66	20.72	5.66	13.19	3.82	20.46	12.14	16.76	38.63	23.25	28.60	16.67
InternVL3-2B	30.20	3.70	16.95	49.29	27.52	38.41	42.37	14.50	28.43	0.63	13.57	7.10	16.81	42.27	22.53	32.23	24.16
Qwen-2-VL-2B	36.44	7.61	22.02	47.81	52.24	50.02	55.70	9.82	32.76	1.13	18.86	9.99	27.04	61.54	23.04	26.08	30.13
Qwen-2.5-VL-3B	36.03	22.43	29.23	52.68	42.35	47.51	44.61	11.89	28.25	10.94	34.16	22.55	20.75	53.69	21.87	38.05	32.31
Open-Source Large Language Models (7B+)																	
Qwen-2-VL-7B	27.58	33.92	30.75	65.10	48.47	56.79	17.60	16.73	17.17	22.59	38.88	30.74	30.17	48.23	28.44	31.15	34.02
Qwen-2.5-VL-7B	44.82	36.91	40.86	76.28	44.34	60.31	36.67	26.86	31.77	20.72	42.43	31.57	29.81	77.95	27.08	34.29	41.42
InternLM-XComposer-2-7B	32.23	27.29	29.76	46.82	52.71	49.77	50.62	28.59	39.61	4.80	40.78	22.79	21.04	71.54	22.55	33.04	35.87
InternLM-XComposer-2.5-7B	47.78	25.07	36.43	60.03	49.57	54.80	50.80	15.84	33.32	7.25	39.14	23.19	28.31	51.07	22.74	31.82	36.07
InternLM-XComposer-2.5-7B-Reward	39.78	37.84	38.81	64.09	55.72	59.90	36.85	21.57	29.21	22.19	11.83	17.01	27.14	65.94	26.94	31.63	36.65
GLM-4V-9B	38.52	35.67	37.09	58.72	56.25	57.49	48.31	15.91	32.11	28.81	22.55	25.68	36.50	68.01	25.31	37.62	39.03
Llama-3.2-11B-Vision	26.81	20.90	31.91	27.50	54.71	49.17	32.56	2.91	25.80	0.20	14.88	15.60	28.67	53.52	39.31	32.02	34.50
Llama-3.2-11B-Vision-Instruct	37.36	31.28	34.32	58.56	55.23	56.90	43.31	14.43	28.87	15.15	38.83	26.99	37.73	54.52	20.67	28.70	36.43
InternVL3-14B	45.21	53.27	49.24	65.04	51.85	58.44	47.03	31.49	39.26	31.99	55.88	43.93	36.89	78.54	23.67	36.90	46.79
Closed-Source Large Language Models (API)																	
Claude-3.7-Sonnet	35.25	0.00	17.63	61.95	61.26	61.60	45.09	19.43	32.26	33.92	70.62	52.27	32.50	63.91	18.55	20.03	39.14
GLM-4V-Plus	44.93	42.58	43.76	70.62	60.54	65.58	43.14	28.75	35.94	33.91	62.52	48.22	40.48	64.38	30.37	36.78	47.03
Doubao-1.5-Vision-Pro	44.12	48.20	46.16	82.10	67.87	74.98	53.85	24.35	39.10	33.57	41.81	37.69	44.00	78.41	17.59	30.78	47.79
GPT-4o-Mini	39.32	49.37	44.35	59.78	80.91	70.34	53.86	33.76	43.81	29.88	49.03	39.46	51.02	64.08	30.08	31.19	48.14
GPT-4o	54.37	60.15	57.26	65.09	68.00	66.54	47.24	28.24	37.74	35.68	58.52	47.10	48.30	75.84	27.82	31.25	50.57
Qwen-VL-Max	40.89	56.79	48.84	64.39	68.72	66.56	49.84	36.58	43.21	35.06	64.28	49.67	37.19	71.50	34.25	36.85	50.29
Gemini-2.5-Pro	52.70	59.13	55.92	77.46	73.15	75.30	47.27	35.12	41.20	34.51	47.17	40.84	61.24	86.52	28.87	38.55	53.43
DinoCompanion	**54.63**	**60.40**	**57.51**	**75.23**	**70.75**	**72.99**	**57.87**	**46.50**	**52.19**	**59.47**	**58.17**	**58.82**	**37.58**	**69.73**	**36.73**	**45.79**	**57.15**

Table 5: Ablation of Carpo components on AttachSecure (higher is better).

Variant	Removed	Avg. Score	Observation
w/o Risk Score	$\lambda = 0$ (no risk penalty)	60.42 (+3.27)	Quality up, risk detection collapses
w/o Uncertainty $\lambda(u)$	Constant λ	55.03	More high-risk misclassifications
w/o KL Constraint	$\beta \to \infty$ (no KL)	52.28	Fluency up, persona drift severe

Table 6: Performance drop after removing individual sensory channels.

Channel Removed	ER-Recog	SB-Effect	CP-Consist	Avg.
Vision	−19.8	−5.6	≈	−8.2
Speech	−8.1	−4.3	≈	−4.9

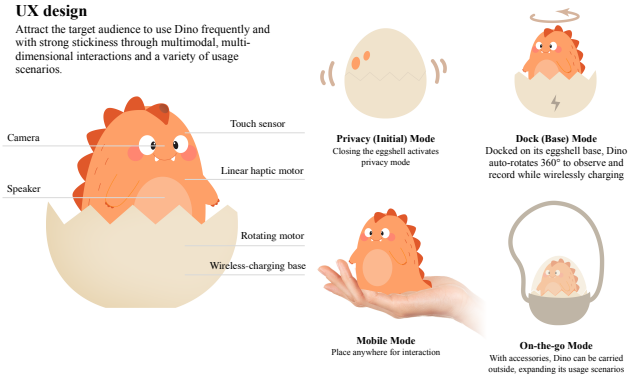


Figure 5: User experience design of DinoCompanion.

The 3D-printed modular shell houses a camera, speaker, touch sensor, and dual motors (linear haptic and rotating) in a compact wireless-charging design. Four interaction modes support diverse use cases: (i) privacy mode via shell closure, (ii) 360° observation when docked, (iii) mobile placement, and (iv) portable accessories for outdoor scenarios.

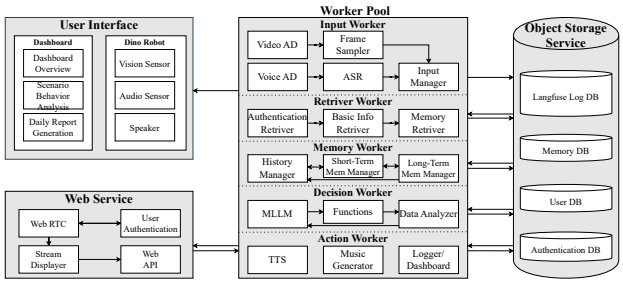


Figure 6: System architecture of DinoCompanion.

System Architecture. Figure 6 presents DinoCompanion's GPU-accelerated backend architecture. The multimodal input pipeline fuses visual and audio streams before passing to the LLM-based decision module, which integrates persistent memory and CARPO-balanced response generation. Real-time outputs leverage TTS and adaptive music generation, while WebRTC enables secure caregiver monitoring through the web dashboard. This modular design ensures both child safety and developmental appropriateness across all interactions.

7 Conclusion

We introduce DinoCompanion, a multimodal robot grounded in attachment theory, aimed at enhancing emotionally responsive child-AI interactions. By integrating developmental psychology and multimodal capabilities, we address key challenges in engagement and emotional safety. The CARPO framework ensures emotional alignment, and the AttachSecure-Bench benchmark enables effective evaluation. DinoCompanion outperforms existing models in attachment-related competencies, paving the way for safer, developmentally informed AI companions for children.

GenAI Usage Disclosure

We know that the ACM's Authorship Policy requires full disclosure of all use of generative AI tools in all stages of the research (including the code and data) and the writing. No GenAI tools were used in any stage of the research, nor in the writing.

References

[1] Nida Itrat Abbasi, Micol Spitale, Joanna Anderson, Tamsin Ford, Peter B Jones, and Hatice Gunes. 2022. Can robots help in the evaluation of mental wellbeing in children? an empirical study. In *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 1459–1466.

[2] Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus* 3, 7 (2024), pgae245.

[3] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 34–44.

[4] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Do LLMs Exhibit Human-Like Reasoning? Evaluating Theory of Mind in LLMs for Open-Ended Responses. *arXiv preprint arXiv:2406.05659* (2024).

[5] Anthropic. 2025. Claude 3.7 Sonnet. *Anthropic News* (February 2025). https://www.anthropic.com/news/claude-3-7-sonnet

[6] Sinem Aslan, Lenitra M Durham, Nese Alyuz, Eda Okur, Sangita Sharma, Celal Savur, and Lama Nachman. 2024. Immersive multi-modal pedagogical conversational artificial intelligence for early childhood education: An exploratory case study in the wild. *Computers and Education: Artificial Intelligence* 6 (2024), 100220.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[9] Priyank Bhutani, Manu Gupta, Gagan Bajaj, Ramesh Chandra Deka, Siddhartha Sankar Satapathy, and Suvendra Kumar Ray. 2024. Is the screen time duration affecting children's language development?-A scoping review. *Clinical Epidemiology and Global Health* 25 (2024), 101457.

[10] Valérie Brauchli, Fabio Sticca, Peter Edelsbrunner, Agnes von Wyl, and Patricia Lannen. 2024. Are screen media the new pacifiers? The role of parenting stress and parental attitudes for children's screen time in early childhood. *Computers in Human Behavior* 152 (2024), 108057.

[11] Inge Bretherton. 2013. The origins of attachment theory: John Bowlby and Mary Ainsworth. In *Attachment theory*. Routledge, 45–84.

[12] Mary E Brushe, Dandara G Haag, Edward C Melhuish, Sheena Reilly, and Tess Gregory. 2024. Screen time and parent-child talk when children are aged 12 to 36 months. *JAMA pediatrics* 178, 4 (2024), 369–375.

[13] Nicholas Buttrick. 2024. Studying large language models as compression algorithms for human culture. *Trends in cognitive sciences* 28, 3 (2024), 187–189.

[14] ByteDance. 2024. Doubao Model Series. https://www.doubao.com/. Accessed: 2025-05-14.

[15] Alan Carr, Laura Finneran, Christine Boyd, Claire Shirey, Ciaran Canning, Owen Stafford, James Lyons, Katie Cullen, Cian Prendergast, Chris Corbett, et al. 2024. The evidence-base for positive psychology interventions: a mega-analysis of meta-analyses. *The Journal of Positive Psychology* 19, 2 (2024), 191–205.

[16] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[17] Ching-Huei Chen and Ching-Ling Chang. 2024. Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies* 29, 14 (2024), 18621–18642.

[18] Liuqing Chen, Shuhong Xiao, Yunnong Chen, Yaxuan Song, Ruoyu Wu, and Lingyun Sun. 2024. ChatScratch: An AI-Augmented System Toward Autonomous Visual Programming Learning for Children Aged 6-12. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[19] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196* (2024).

[20] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. CogBench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225* (2024).

[21] OpenCompass Contributors. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

[22] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[23] Or Dagan, Carlo Schuengel, Marije L Verhage, Marinus H van IJzendoorn, Abraham Sagi-Schwartz, Sheri Madigan, Robbie Duschinsky, Glenn I Roisman, Kristin Bernard, Marian Bakermans-Kranenburg, et al. 2021. Configurations of mother-child and father-child attachment as predictors of internalizing and externalizing behavioral problems: An individual participant data (IPD) meta-analysis. *New Directions for Child and Adolescent Development* 2021, 180 (2021), 67–94.

[24] Pierre Dewitte. 2024. Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review* 54 (2024), 106019.

[25] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420* (2024).

[26] Zohar Elyoseph, Tamar Gur, Yuval Haber, Tomer Simon, Tal Angert, Yuval Navon, Amir Tal, and Oren Asman. 2024. An ethical perspective on the democratization of mental health with generative AI. *JMIR Mental Health* 11 (2024), e58011.

[27] David Estévez, María-José Terrón-López, Paloma J Velasco-Quintana, Rosa-María Rodríguez-Jiménez, and Valle Álvarez-Manzano. 2021. A case study of a robot-assisted speech therapy for children with language disorders. *Sustainability* 13, 5 (2021), 2771.

[28] Chiara Filippini, Edoardo Spadolini, Daniela Cardone, Domenico Bianchi, Maurizio Preziuso, Christian Sciarretta, Valentina Del Cimmuto, Davide Lisciani, and Arcangelo Merla. 2021. Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. *International Journal of Social Robotics* 13 (2021), 677–689.

[29] Manuela Gander, Alexander Karabatsiakis, Katharina Nuderscher, Dorothee Bernheim, Cornelia Doyen-Waldecker, and Anna Buchheim. 2022. Secure attachment representation in adolescence buffers heart-rate reactivity in response to attachment-related stressors. *Frontiers in human neuroscience* 16 (2022), 806987.

[30] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793

[31] Brian P Godor, Frank CP van der Horst, and Ruth Van der Hallen. 2024. Unravelling the roots of emotional development: Examining the relationships between attachment, resilience and coping in young adolescents. *The Journal of Early Adolescence* 44, 4 (2024), 429–457.

[32] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[33] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[34] Hui-Ru Ho, Edward M Hubbard, and Bilge Mutlu. 2024. " It's Not a Replacement:" Enabling Parent-Robot Collaboration to Support In-Home Learning Experiences of Young Children. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[35] Paul Holmes and Steve Farnfield. 2022. *The Routledge Handbook of Attachment (3 Volume Set)*. Taylor & Francis.

[36] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[37] Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519* (2024).

[38] Sarika Kewalramani, Kelly-Ann Allen, Erin Leif, and Andrea Ng. 2024. A scoping review of the use of robotics technologies for supporting social-emotional

learning in children with autism. *Journal of Autism and Developmental Disorders* 54, 12 (2024), 4481–4495.
[39] Nomisha Kurian. 2024. 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap'in large language models. *Learning, Media and Technology* (2024), 1–14.
[40] Nomisha Kurian. 2025. AI's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care. *Contemporary Issues in Early Childhood* 26, 1 (2025), 132–139.
[41] Soyang Kwon, Bridget Armstrong, Nina Wetoska, and Selin Capan. 2024. Screen time, sociodemographic factors, and psychological well-being among young children. *JAMA network open* 7, 3 (2024), e2354488–e2354488.
[42] Valentina Lucia La Rosa, Alessandra Geraci, Alice Iacono, and Elena Commodari. 2024. Affective touch in preterm infant development: neurobiological mechanisms and implications for child–caregiver attachment and neonatal care. *Children* 11, 11 (2024), 1407.
[43] Amanda Lagerkvist, Matilda Tudor, Jacek Smolicki, Charles M Ess, Jenny Eriksson Lundström, and Maria Rogg. 2024. Body stakes: an existential ethics of care in living with biometrics and AI. *AI & SOCIETY* 39, 1 (2024), 169–181.
[44] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675* (2024).
[45] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A Survey of Multimodel Large Language Models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*. 405–409.
[46] Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253* (2024).
[47] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences* (2024).
[48] André Markus, Jan Pfister, Astrid Carolus, Andreas Hotho, and Carolin Wienrich. 2024. Effects of AI understanding-training on AI literacy, usage, self-determined interactions, and anthropomorphization with voice assistants. *Computers and Education Open* 6 (2024), 100176.
[49] Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one* 19, 3 (2024), e0298522.
[50] Graham Music. 2024. *Nurturing natures: Attachment and children's emotional, sociocultural and brain development*. Routledge.
[51] Nazneen Nazeer, Jenny Parker, Lauren Cross, Sophie Epstein, Jessica Penhallow, Tamsin Newlove-Delgado, Johnny Downs, and Tamsin Ford. 2025. The extent to which child and parent report RCADS, sMFQ, SDQ and child report KIDSCREEN identify the same young people as at risk of mental health conditions. *The British journal of psychiatry: the journal of mental science* (2025), 1.
[52] Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *arXiv preprint arXiv:2403.10779* (2024).
[53] Jessica E Opie, Jennifer E McIntosh, Timothy B Esler, Robbie Duschinsky, Carol George, Allan Schore, Emily J Kothe, Evelyn S Tan, Christopher J Greenwood, and Craig A Olsson. 2021. Early childhood attachment stability and change: A meta-analysis. *Attachment & Human Development* 23, 6 (2021), 897–930.
[54] Ekaterina Pashevich. 2022. Can communication with social robots influence how children develop empathy? Best-evidence synthesis. *AI & SOCIETY* 37, 2 (2022), 579–589.
[55] Abigail E Pine, Mary G Baumann, Gabriella Modugno, and Bruce E Compas. 2024. Parental involvement in adolescent psychological interventions: a meta-analysis. *Clinical Child and Family Psychology Review* 27, 3 (2024), 1–20.
[56] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access* 12 (2024), 26839–26874.
[57] Matan Rubin, Hadar Arnon, Jonathan D Huppert, Anat Perry, et al. 2024. Considering the role of human empathy in AI-driven therapy. *JMIR Mental Health* 11, 1 (2024), e56529.
[58] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416* (2024).
[59] Carlo Schuengel, Marije L Verhage, and Robbie Duschinsky. 2021. Prospecting the attachment research field: A move to the level of engagement. *Attachment & Human Development* 23, 4 (2021), 375–395.
[60] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. Chacha: leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.
[61] Faruk Seyitoğlu and Stanislav Ivanov. 2024. Robots and emotional intelligence: A thematic analysis. *Technology in Society* 77 (2024), 102512.
[62] Deborah Silvis, Jody Clarke-Midura, Jessica F Shumway, Victor R Lee, and Selendra Mullen. 2022. Children caring for robots: Expanding computational thinking frameworks to include a technological ethic of care. *International Journal of Child-Computer Interaction* 33 (2022), 100491.
[63] Luke Stark. 2024. Animation and Artificial Intelligence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1663–1671.
[64] Alexandra R Tabachnick, Yunqi He, Lindsay Zajac, Elizabeth A Carlson, and Mary Dozier. 2022. Secure attachment in infancy predicts context-dependent emotion expression in middle childhood. *Emotion* 22, 2 (2022), 258.
[65] Manyu Tang, Yongcai Chen, et al. 2024. AI and animated character design: efficiency, creativity, interactivity. *The Frontiers of Society, Science and Technology* 6, 1 (2024), 117–123.
[66] Annalisa Tanzilli, Mariagrazia Di Giuseppe, Guido Giovanardi, Tommaso Boldrini, Giorgio Caviglia, Ciro Conversano, and Vittorio Lingiardi. 2021. Mentalization, attachment, and defense mechanisms: a Psychodynamic Diagnostic Manual-2-oriented empirical investigation. *Research in Psychotherapy: Psychopathology, Process, and Outcome* 24, 1 (2021), 531.
[67] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[68] Carolina Villegas-Galaviz and Kirsten Martin. 2024. Moral distance, AI, and the ethics of care. *AI & society* 39, 4 (2024), 1695–1706.
[69] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).
[70] Xizheng Xu, Zhiqiang Liu, Shaoying Gong, and Yunpeng Wu. 2022. The relationship between empathy and attachment in children and adolescents: Three-level meta-analyses. *International Journal of Environmental Research and Public Health* 19, 3 (2022), 1391.
[71] Yijia Yuan. 2024. An empirical study of the efficacy of AI chatbots for English as a foreign language learning in primary education. *Interactive Learning Environments* 32, 10 (2024), 6774–6789.
[72] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. 2025. InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model. In *Findings of ACL*.
[73] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through Child-AI co-creative storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.
[74] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output. *arXiv preprint arXiv:2407.03320* (2024).
[75] Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. 2024. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638* (2024).
[76] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.
[77] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479 [cs.CV] https://arxiv.org/abs/2504.10479